

# Assessing the agreement in retraction indexing across 4 multidisciplinary sources: Crossref, Retraction Watch, Scopus, and Web of Science

Jodi Schneider\*, Jou Lee\*\*, Heng Zheng\*\*, Malik Salami\*\*

*\*jodi@illinois.edu & jschneider@pobox.com*

0000-0002-5098-5667

School of Information Sciences, University of Illinois at Urbana-Champaign, USA

*\*\*joulee2@illinois.edu; zhenghz@illinois.edu; malikos2@illinois.edu*

0000-0001-8927-0370; 0000-0001-5866-7746; 0000-0002-2329-5660

School of Information Sciences, University of Illinois at Urbana-Champaign, USA

Previous research has posited a correlation between poor indexing and inadvertent post-retraction citation. However, to date, there has been limited systematic study of retraction indexing quality: we are aware of one database-wide comparison of PubMed and Web of Science, and multiple smaller studies highlighting indexing problems for items with the same reason for retraction or same field of study. To assess the agreement between multidisciplinary retraction indexes, we create a union list of 49,924 publications with DOIs from the retraction indices of at least one of Crossref, Retraction Watch, Scopus, and Web of Science. Only 1593 (3%) are deemed retracted by the intersection of all four sources. For 10,512 publications (21%), there is disagreement: at least one source deems them retracted while another lacks retraction indexing. Of the items deemed retracted by at least one source, retraction indexing was lacking for 32% covered in Scopus, 7% covered in Crossref, and 4% covered in Web of Science. We manually examined 201 items from the union list and found that 115/201 (57.21%) DOIs were retracted publications while 59 (29.35%) were retraction notices. In future work we plan to use a validated version of this union list to assess the retraction indexing of subject-specific sources.

## 1. Introduction

Retraction has been widely studied in scientometric research, often relying on databases such as PubMed and Web of Science to determine which publications are retracted. Only 5.4% of post-retraction citations in PubMed Central acknowledged that the paper they were citing was retracted (Hsiao & Schneider, 2021), and a case study posited a correlation between poor indexing and inadvertent post-retraction citation (Schneider et al., 2020).

Many retracted papers are not marked as retracted on publisher and aggregator sites (Badreldin et al., 2020; Decullier & Maisonneuve, 2018). Retraction status is inconsistently displayed across a wide range of sources, including publisher sites (Dal-Ré & Ayuso, 2020; Suelzer et al., 2021), search engines (Genot & Olsson, 2021), scholarly databases (Mine, 2019; Proescholdt & Schneider, 2020; Schneider et al., 2020; Suelzer et al., 2021), and other websites (Frampton et al., 2021; Mine, 2019). Retraction status can be difficult to discover or confirm (Schneider et al., 2020), because most databases do not treat the retracted paper and retraction and retraction notice “as an integrated entity” (Wang, 2023).

Retraction indexing may also be lacking in some cases. For example, Proescholdt & Schneider (2020) found thousands of examples of apparently retracted papers that were not indexed as such, whose titles starting with "RETRACTED:" or a cognate phrase. Early retractions might also pose challenges: many were issued in non-citable ways such as “tip-in” notices (Snodgrass & Pfeifer,

1992), which did not meet PubMed indexing standards (Kotzin & Schuyler, 1989) and would be missed by retraction indexing. Other studies discovered indexing issues in both document titles and the linking of retracted publications and retraction notices (Schmidt, 2018; Suelzer et al., 2021).

However, to date, there has been limited systematic study of retraction indexing quality: we are aware of one database-wide comparison of PubMed and Web of Science (Schmidt, 2018), and multiple smaller studies highlighting database indexing problems for items with the same reason for retraction (e.g., Malički et al., 2019) or same field of study (e.g., Bakker & Riegelman, 2018; Dal-Ré & Ayuso, 2020; among many others). An analysis of PubMed's duplicate publication index in 2013 found 48% (12/25) of retracted publications (identified by publisher notices) did not show retraction status correctly for duplicate publications, and these problems persisted after authors contacted PubMed and editors during a 5-year follow-up period (Malički et al., 2019). 38% of mental health articles and 4% of genetics articles marked as retracted in Retraction Watch were not indexed as retracted in PubMed (Bakker & Riegelman, 2018; Dal-Ré & Ayuso, 2020). An analysis of 144 retracted articles in mental health found that only 7% (10/144) of retracted items were marked as such across a variety of publisher sites and database records (i.e., EBSCO databases, MEDLINE and PsycINFO via Ovid, PubMed, Scopus, Web of Science), and of those, the majority only indicated the retraction in one place (Bakker & Riegelman, 2018).

While it is known that retraction indexes are incomplete, there has been no systematic assessment of the extent to which retraction metadata agrees in multidisciplinary databases. This study fills that gap.

## **2. Goals and Research Questions**

We construct a union list of all DOIs indexed as retracted publications in at least one of four multidisciplinary sources: Crossref, Retraction Watch, Scopus, and Web of Science. We check the extent to which each source agrees with the union list, restricting to each source's coverage.

Our specific research questions are:

- RQ1: How many DOIs are indexed as retracted publications in each of Crossref, Retraction Watch, Scopus, and Web of Science? Overall, how many DOIs are indexed as retracted publications in at least one source?
- RQ2: How much agreement does each source have with the union list, restricting to its coverage?
- RQ3: Does the level of agreement in DOIs indexed as retracted publications vary by field, publication year, or retraction year?
- RQ4: For a sample of DOIs with less than 100% agreement in retraction indexing, does the publisher's website indicate that they are retracted publications?

## **3. Methods and Data**

*3.1. Methods and Data for RQ1: How many DOIs are indexed as retracted publications in each of Crossref, Retraction Watch, Scopus, and Web of Science? Overall, how many DOIs are indexed as retracted publications in at least one source?*

To address RQ1, we create a list of DOIs that are indexed as retracted publications in one or more of our sources. To do this, we extract metadata about retracted publications as shown in Table 1.

After retrieving DOIs indexed as retracted publications, we deduplicate metadata within each data source, removing duplicate items with the same DOI. For ease of matching, we also remove items without DOI. Then we combine metadata across the four sources. Each DOI is annotated with a list of the sources that indexed it as a retracted publication, which we call `rp_indexed_in`. We do not seek to retrieve publications indexed as errata or correction because according to the Committee on Publishing Ethics (COPE) (2019), retractions should be distinguished from other types of correction or comment.

Table 1. Retracted publications identified from multidisciplinary sources.

Source	Search Query	Query Results Retrieved <sup>1</sup>	Search Date	Top Categories (as categorized by source)
Crossref	Update_type=( 'retraction', 'Retraction', 'retracion', 'retration', 'partial_retraction', 'withdrawal','removal')	14,745	2023-04-05	General Medicine (1738); Pharmacology (medical) (1315); Multidisciplinary (883); General Computer Science (426); General Environmental Science (385); Biochemistry (385)
Retraction Watch	All results	39,301	2023-03-27	((BLS) Biology - Cancer;(BLS) Biology - Cellular;(BLS) Genetics;(838)  (B/T) Computer Science;(B/T) Technology; (719)  (B/T) Computer Science; (674)
Scopus <sup>2</sup>	DOCTYPE("tb") <sup>3</sup>	21,515	2023-04-05	Computer Science (6,911) Engineering (5,887) Medicine (3,908) Biochemistry, Genetics and Molecular Biology (2,935) Business, Management and Accounting (2,884) Physics and Astronomy (2,078)
Web of Science (WOS) all collections	DT="Retracted Publication"	16,434	2023-04-05	Biochemistry Molecular Biology (7,920) Genetics Heredity (5,796) Cell Biology (5,495) Pharmacology Pharmacy (5,010) Oncology (4,225) Immunology (2,810)

<sup>1</sup> As retrieved from each data source, before deduplication and before checking for DOIs.

<sup>2</sup> This data was downloaded from Scopus API on April 5, 2023 via <http://www.scopus.com>

<sup>3</sup> Retracted publications can be retrieved with the tb document type in Scopus: [https://service.elsevier.com/app/answers/detail/a\\_id/11236/supporthub/scopus/](https://service.elsevier.com/app/answers/detail/a_id/11236/supporthub/scopus/)

*3.2. Methods and Data for RQ2: How much agreement does each source have with the union list, restricting to its coverage?*

An item might not be found in a given source on a given search date, because either: the item was not covered by the source; or, the item was covered but is not indexed as a retracted publication in that source. For a given DOI, we poll each source that it is not "rp\_indexed\_in" (using results from RQ1), to see whether the DOI is "covered\_in" the source. We use APIs for Crossref, Scopus, and Web of Science; for Retraction Watch, there is nothing to check because our database dump only covers retracted publications.

In calculating agreement, we will consider a source to agree if it indexes as retracted a publication that is deemed retracted by any one of our sources (including just itself).

Considering the coverage, we quantify the extent of the agreement in retraction indexing for each source:

$$\text{RetractionIndexingAgreement\_SOURCE} = \frac{\text{Number of DOIs rp\_indexed\_in(SOURCE)}}{\text{Number of DOIs covered\_in(SOURCE)}}$$

*3.3. Methods and Data for RQ3: Does the level of agreement in DOIs indexed as retracted publications vary by field, publication year, or retraction year?*

Analogous to the RetractionIndexingAgreement\_SOURCE above, we also quantify the extent of the agreement in retraction indexing for each DOI:

$$\text{RetractionIndexingAgreement\_DOI} = \frac{\text{Number of sources the DOI is rp\_indexed\_in}}{\text{Number of sources the DOI is covered\_in}}$$

We then analyze the RetractionIndexingAgreement\_DOI across field, publication year, and retraction year.

We (JL, JS, MS) categorize DOIs based on their venue (conference or journal). First we clean venue titles: we remove all digits, all hyphens ('-'), all extra whitespace, and all positional words (e.g., 22nd, 11th) as well as 'the' at the start of titles. With OpenRefine<sup>4</sup>, we reconcile differences in acronym position (e.g., 'international renewable energy congress irec' vs. 'irec international renewable energy congress') and non-Latin characters, e.g., 'almasäq' and 'almasāq'. We then compare venue titles against the Scopus source list<sup>5</sup>. We categorize venues according to Scopus's categorization when available, as one or more of Health Science, Life Science, Physical Science, Social Science, or General. The Scopus source list contained 62% (4926/7907) of the venue titles, which covered 60% (30,002/49,924) of the DOIs.

---

<sup>4</sup> <https://openrefine.org>

<sup>5</sup> <https://www.elsevier.com/?a=91122> ; in July 2023 this retrieved the list updated through June 2023, extlistJune2023.xlsx

For the remaining venues not in the Scopus source list (2981/7907 or 37%), we use a multi-step process to assign our own categories. First, we identify content words associated with each category, using the Scopus source list as a resource. We do this by first removing stopwords<sup>6</sup> from venue titles and then using Yet Another Keyword Extractor<sup>7</sup>.

Using this list of content words, we compare with the remaining uncategorized venue titles. We assign a venue title to a category (potentially multiple categories) when there is a match in content words. We iteratively review uncategorized venue titles to manually curate additional lists of content words. Any venue title that remained unclassified after the several rounds of iteration is placed in a new category which we call the “uncategorized” category, which is separate from the “General” category above. Uncategorized venue titles include university-specific journals (e.g., “Bulletin of Mgimo University”), titles relying on underspecified terms (e.g., “The Journal of ECT”), terms with multiple potential meanings (e.g., “Challenge”), and general titles with no content phrases (e.g., “Science International”, “Colloquium-Journal”, “Contexto Internacional”). We left as uncategorized venue titles that could not be classified with our content keyword principle, including those with person names that are not strictly content words (e.g., “Einstein”, “Antonie Van Leeuwenhoek”); that required multi-word content phrases (e.g., “Journal of Frontier Studies”, “Substance Use & Misuse”); or contain non-English words that did not yield useful automatic translations (e.g., “Al-MasÄq”).

#### *3.4. Methods and Data for RQ4: For a sample of DOIs with less than 100% agreement in retraction indexing, does the publisher’s website indicate that they are retracted publications?*

We (HZ, JS) examine a sample of 201 DOIs from our union list that are covered in multiple sources that disagree on their retraction indexing (e.g., `RetractionIndexingAgreement_DOI < 100%`), to check: Does the publisher’s website indicate that they are retracted publications?

To select the sample, we first group DOIs using the `RetractionIndexingAgreement_DOI` score as calculated from RQ2 and then select items from each group. We keep other aspects as diverse as feasible, particularly the venue title. We overselect DOIs with certain features: retraction year earlier than the publication year (especially more than 1 year earlier), having a PubMed ID (since PubMed retraction status is public domain data freely available for reuse), or no retraction year in our data.

---

<sup>6</sup> For stopwords, we use both publicly available ISO Stopwords in English, French, German, Latin and Russian <https://github.com/stopwords-iso/stopwords-iso> and our own curated stopword list to handle non-content words like “conference” and “journal”; see <https://github.com/infoqualitylab/retraction-indexing-agreement/blob/main/stopwords.txt> which is also available as `stopwords.txt` in our formal data deposit: [https://doi.org/10.13012/B2IDB-8847584\\_V2](https://doi.org/10.13012/B2IDB-8847584_V2)

<sup>7</sup> <https://pypi.org/project/yake/>

## 4. Results

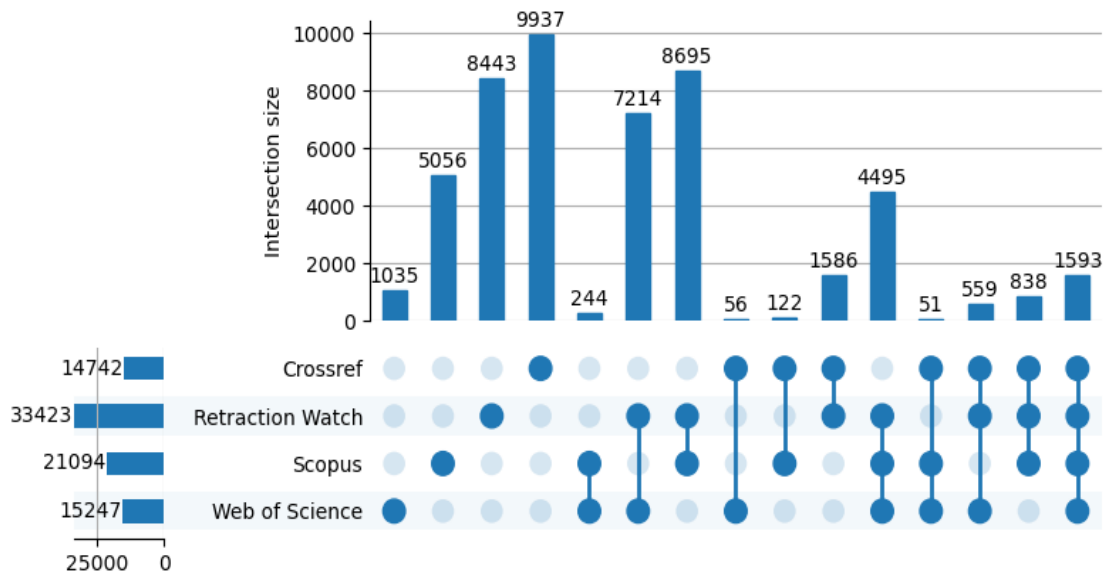
*4.1. Results for RQ1: How many DOIs are indexed as retracted publications in each of Crossref, Retraction Watch, Scopus, and Web of Science? Overall, how many DOIs are indexed as retracted publications in at least one source?*

Our union list has 49,924 unique DOIs that are indexed as a retracted publication by one or more of Crossref, Retraction Watch, Scopus, and Web of Science. As shown in Table 2, these were consolidated and merged from the 91,995 records retrieved.

Table 2. After deduplication and checking for DOIs, we get a merged list of 49,924 unique records with DOI.

Source	Query results retrieved	Records with DOI	Records without DOI removed	Duplicate records removed
Crossref	14,745	14,742	0	3
Retraction Watch	39,301	33,423	5828	50
Scopus	21,515	21,094	49	372
Web of Science	16,434	15,247	1126	61
<b>Total</b>	<b>91,995</b>	<b>84,506</b>	<b>7003</b>	<b>486</b>
<b>Total (Unique)</b>		<b>49,924</b>		

Figure 1: The 49,924 unique DOIs were retrieved as retracted publications from 1, 2, 3, or 4 different sources. In this Upset diagram, following the conventions of (Lex et al., 2014), circles represent sources and lines between the circles indicate the overlap between sources, i.e., the number of DOIs retrieved as retracted publications from the given combination of sources.



As shown in Figure 1, the 49,924 unique DOIs were retrieved as retracted publications from 1, 2, 3, or 4 different sources. Every combination of sources can be read off this Upset diagram (Lex et

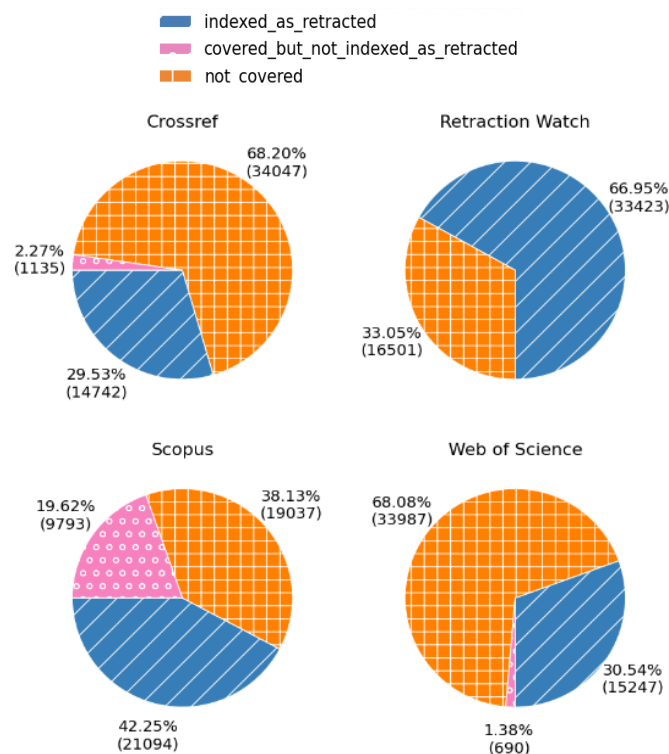
al., 2014); here overlap between sources indicates the number of DOIs indexed as retracted in the given combination of sources.

Among the 49,924 unique DOIs, only 1593 (3%) were found in all four sources, with a total of 24471 (49%) purportedly retracted publications found in only a single source: 9937 (20%) in Crossref, 8443 (17%) in Retraction Watch, 5056 (10%) in Scopus and 1035 (2%) in Web of Science.

#### 4.2. Results for RQ2: How much agreement does each source have with the union list, restricting to its coverage?

The RetractionIndexingAgreement\_SOURCE indicates the percentage of covered items, shared with the union dataset, that are indexed as retracted. Agreement is 100% for Retraction Watch, which only provided retracted publications; 95.67% for Web of Science; 92.85% for Crossref; and 62.29% in Scopus. Coverage differs for each database, and Figures 2 and 3 compare the number of DOIs from our union list that are indexed as retracted in a source (blue) with those covered but not indexed as retracted (pink) and not covered (orange) in that source. Coverage was checked April 9, 2023 with the Crossref API, Scopus API<sup>8</sup>, and the Web of Science API.<sup>9</sup>

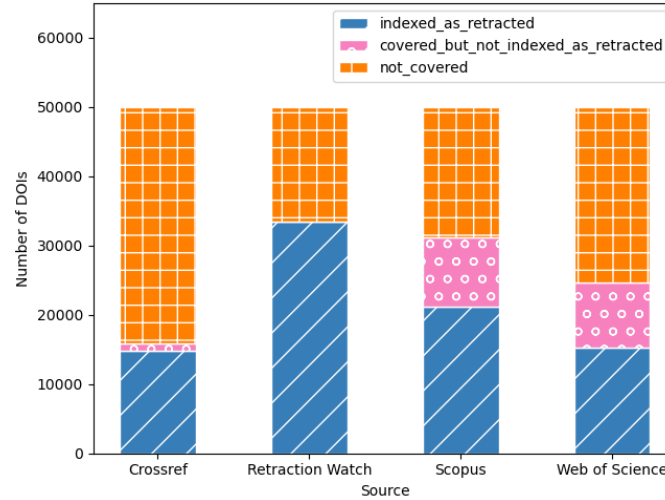
Figure 2: The proportion of our 49,924 DOIs that are: not covered; covered but not indexed as retracted; and indexed as retracted in each of Crossref, Retraction Watch, Scopus, and Web of Science.



<sup>8</sup> via <http://api.elsevier.com> and <http://www.scopus.com>

<sup>9</sup> No separate search is needed in Retraction Watch since it only covers items it deems retracted.

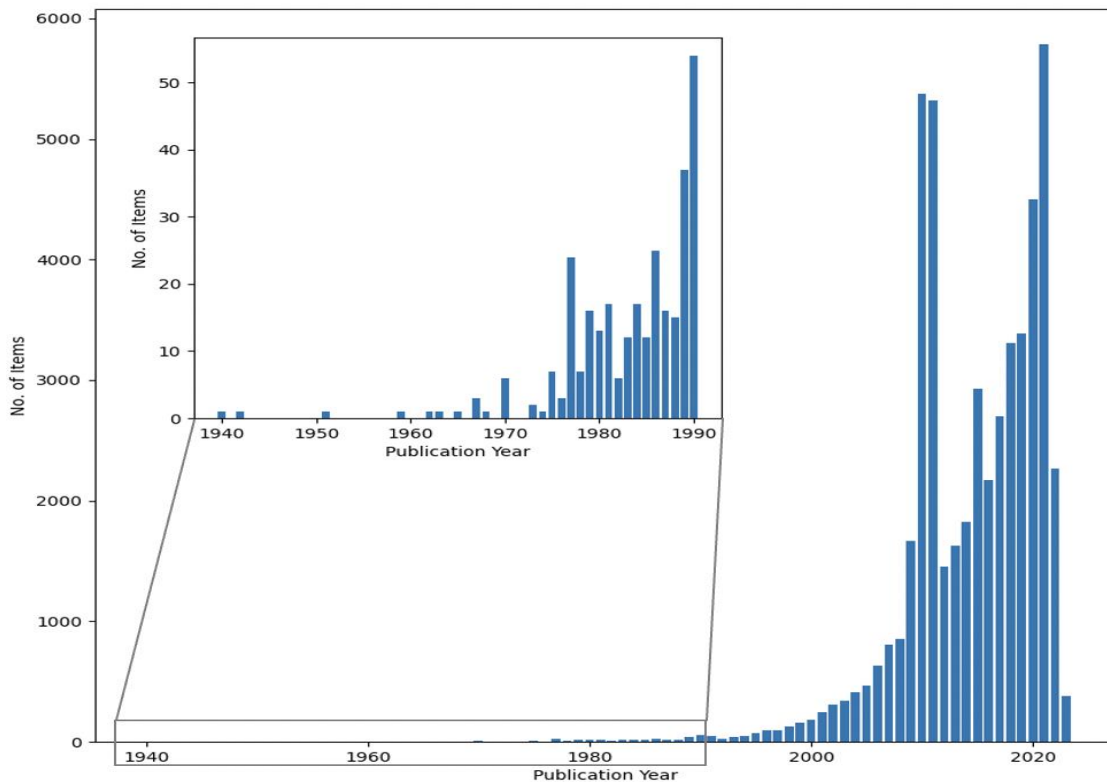
Figure 3: Number of records that are covered but not indexed as retracted; covered and indexed as retracted in each source.



#### 4.3. Results for RQ3: Does the level of agreement in DOIs indexed as retracted publications vary by field, publication year, or retraction year?

While publication years range from 1940 to 2023 (Figure 4), interestingly, the first disagreement in for DOIs in our union list is in publication year 2016: about 570 DOIs were covered but not indexed as retracted in some source. The highest disagreement of over 2000 DOIs was recorded in 2019.

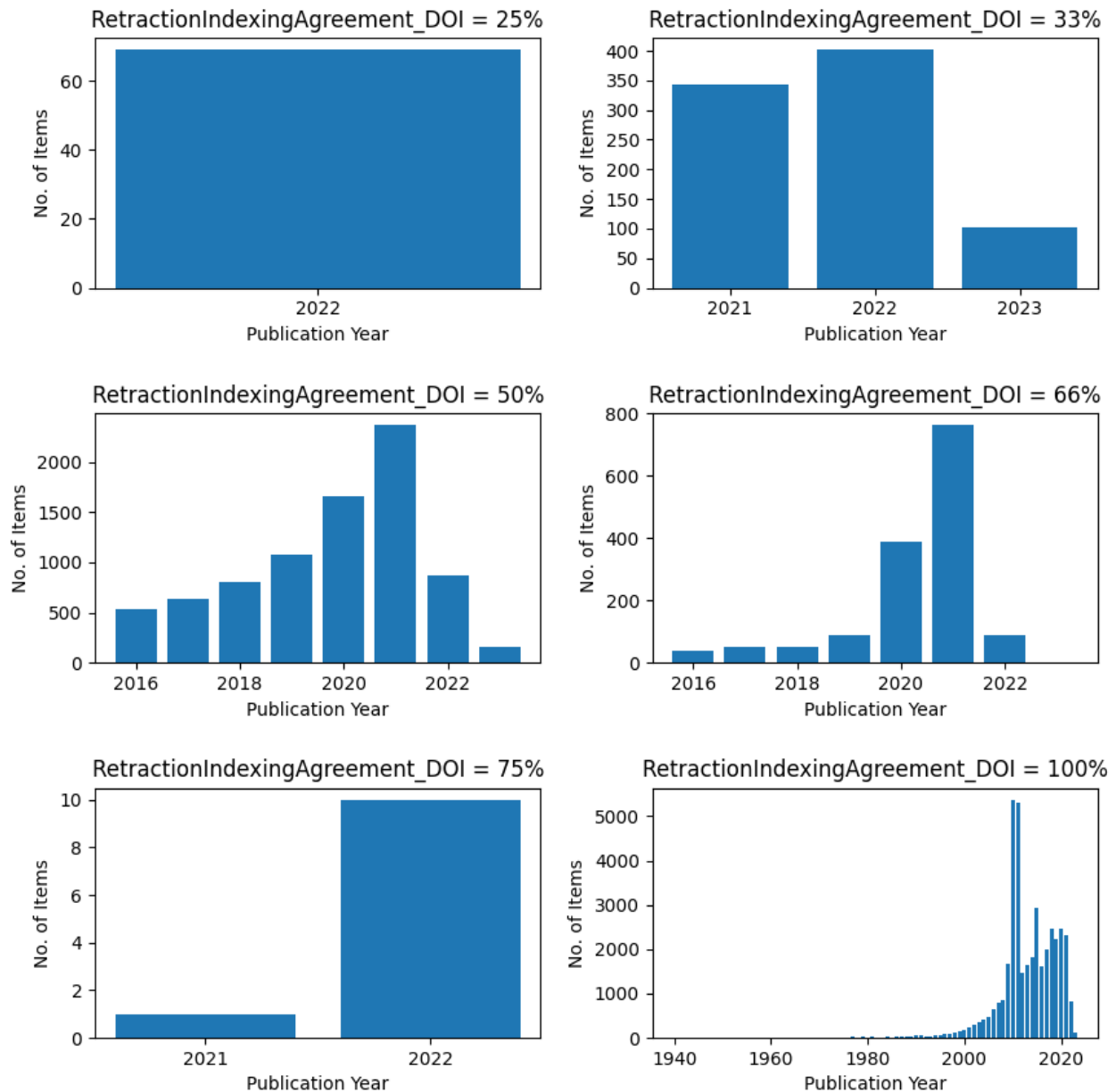
Figure 4: Publication year distribution for our 49,924 DOIs.





The publication year distribution varies by RetractionIndexingAgreement\_DOI, and as shown in Figure 5, agreement of 50% and 66% is found from 2016 forward. By contrast, 25% agreement is found only in publications from 2022; 33% agreement is found only in publications from 2021 to 2023; and 75% agreement is found mostly in publications from 2022 with some from 2021.

Figure 5: Publication year distribution for each RetractionIndexingAgreement\_DOI score.



The retraction year distribution (Figure 6) is roughly similar to the publication year distribution. We have the retraction year for 43,584 DOIs (87%). All DOIs from Retraction Watch include a retraction year. Currently we lack retraction year for 6340 items, those we found only in Scopus (4869, 9.75%), only in WoS (1035, 2.07%), only in Crossref (1, 0%), both Scopus and WoS (245, 0.49%), only in Crossref and Scopus (154, 0.31%), and Crossref, Scopus, Web of Science (36, 0.07%).

Figure 6: Retraction year distribution for each RetractionIndexingAgreement\_DOI score. Limited to the 43,584 (87%) DOIs with retraction year in our records.

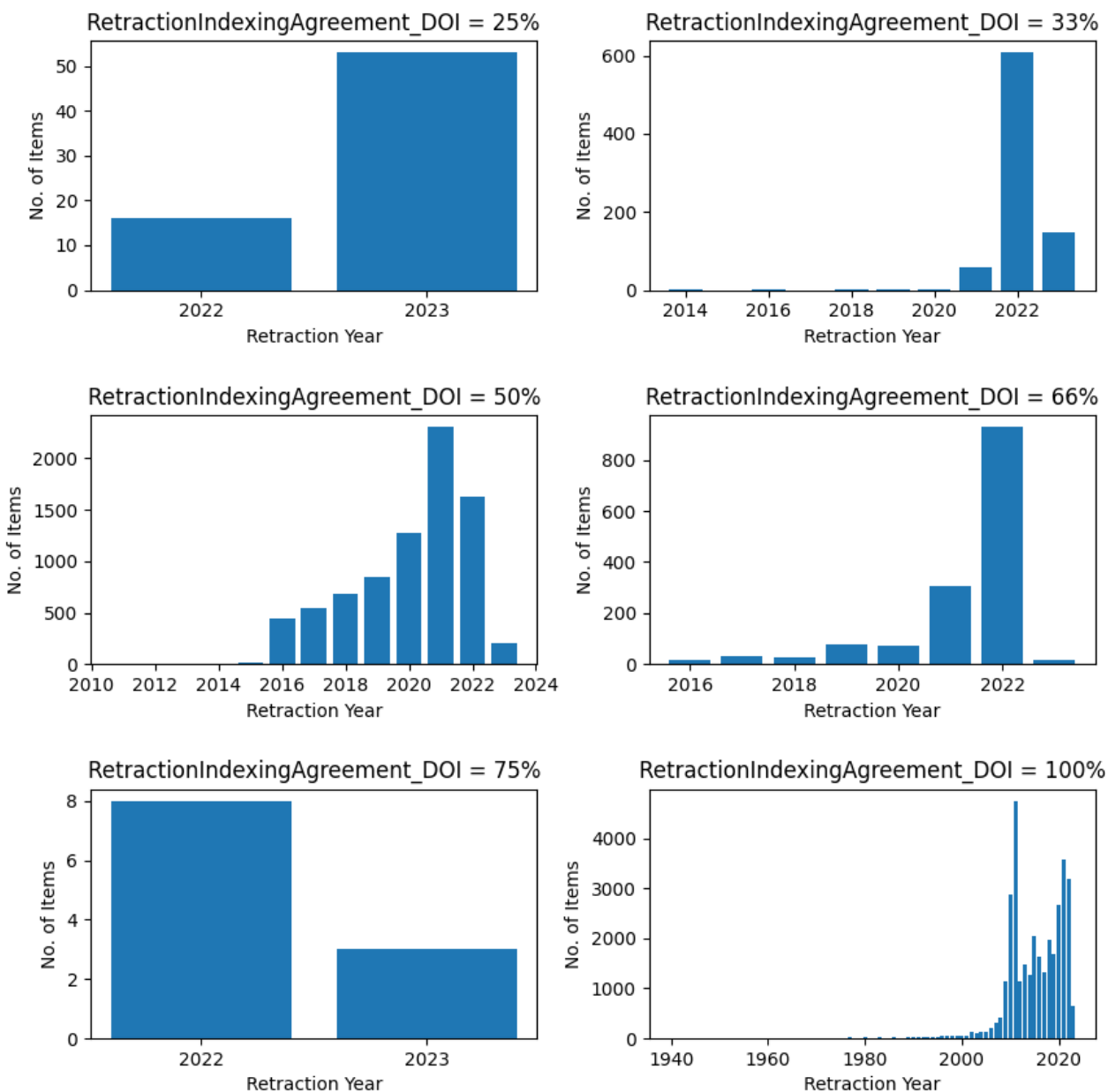
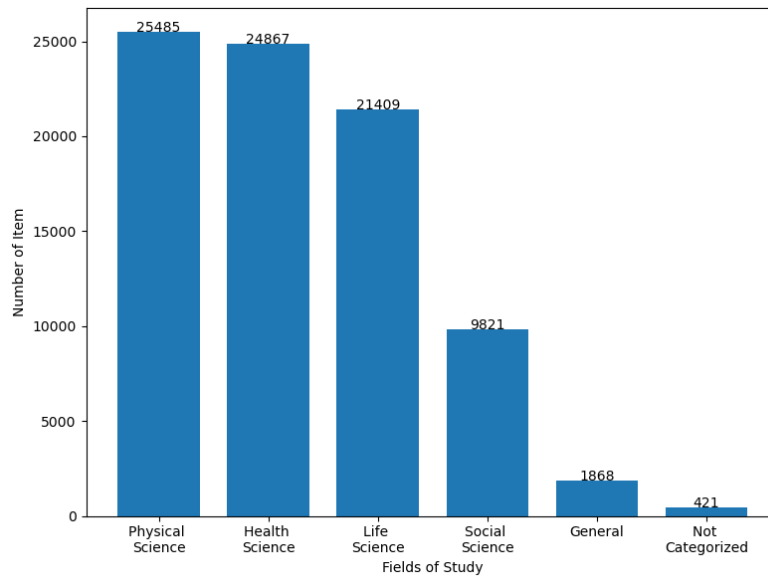


Figure 7 shows the prevalence of Life Science, and to a lesser extent Physical Science, and Health Science DOIs.

Figure 7: Field categorization of the 49,924 DOIs.



*4.4. Results for RQ4: For a sample of DOIs with less than 100% agreement in retraction indexing, does the publisher's website indicate that they are retracted publications?*

In the union list, 10,512 DOIs had RetractionIndexingAgreement\_DOI less than 100%; we sampled 201 (201/10,512=1.9%) of these DOIs for manual review.

Table 3. Categorization of 201 DOIs we manually checked.

Number of DOIs	Percentage	Description
115	57.12%	Retracted publication (including withdrawn or removed articles) <sup>10</sup>
59	29.35%	Retraction notice
14	6.97%	Non-retracted publication that has a correction
11	5.47%	Retraction-related publication
2	1.00%	No sign of retraction

We confirmed that 115/201 (57.12%) DOIs were retracted publications (including withdrawn or removed) as shown in Table 3. The most common indexing error was retraction notices 59/201 (29.35%).<sup>11</sup> We group in “Retraction-related publication” expressions of concern, temporary

<sup>10</sup> Fully distinguishing these categories is difficult because publishers may leave in place the full-text of article as described as withdrawn or take down the full-text of article they describe as retracted. Of the 201 DOIs we checked, 87/201 (43.28%) were retracted articles, 24/201 (11.94%) were withdrawn articles, and 4/201 (1.99%) were removed articles in our judgement.

<sup>11</sup> We counted as retracted publications 12/201 (5.97%) DOIs that are shared by both the retracted article and its retraction notice.

removals, and retracted and republished articles; removed or purportedly retracted publications whose retraction notice we could not immediately locate; and a few retraction-related publications, such as publications whose duplicates had been removed/retracted. In the supplement<sup>12</sup>, Table S1 provides a field breakdown for Table 3 while Table S2 compares the field distribution of the items we manually checked, the set we sampled from, and the entire union list.

## 5. Discussion and Conclusions

We created a union list of DOIs indexed as retracted in one or more of Crossref, Retraction Watch, Scopus, and Web of Science. Among the 49,924 unique DOIs, only 1593 (3%) were found in all four sources, with a total of 24,471 (49%) purportedly retracted publications found in only one source. Agreement with the union list, taking coverage into consideration, is 100% for Retraction Watch, which only provided retracted publications; 95.67% for Web of Science; 92.85% for Crossref; and 62.29% in Scopus. The retraction year and publication year distribution are roughly similar, with disagreements starting in 2016 and most disagreements in publications from 2021 forward with retraction year of 2022 or later.

### 5.1 Limitations

We only examined a very small number of articles (201) manually. Some DOIs indexed as retracted publications were not, in fact, retracted, withdrawn, or removed; many were retraction notices.

We removed 7003 records that had no DOI. We estimate we have lost information about 8-12% of our records (Range is  $5928 - 1126 - 49 = 4753$  to  $7003 / [7003 + 49924]$ ) that have no DOI. Among our sources, Retraction Watch had 5928 records without DOIs; Scopus 49 records without DOIs; and Web of Science 1126 records without DOIs as shown in Table 2.

In calculating agreement metrics, we have a choice in how to handle the DOIs that were uniquely contributed by each source. We have defined our agreement metric to focus on the absence of DOIs contributed by any source (including the source under examination). A stronger metric would consider the presence of unique items a disagreement.

### 5.2 Discussion and Future Work


Disagreement in retraction indexes seems largely to be due to two types of errors: retracted publications with DOIs missing retraction indexing in a source that covers them; and misindexing of DOIs, especially retraction notices and corrigenda.

In the future we would like to better understand how the metadata flows between sources. Multiple types of problems in the metadata flow seem likely. For example, in examining the data we also find discrepancies between publisher websites and metadata; for example, Figure 8 shows that an item with a retraction year discrepancy: 2022 from the publisher website versus 2019 from the Crossref metadata.

---



<sup>12</sup> Our supplement is available at <https://hdl.handle.net/2142/120613>


Figure 8: Discrepancies in data for DOI:10.1016/j.yexmp.2018.12.005 as of April 15, 2023.  
 Left, publisher page from ScienceDirect <https://doi.org/10.1016/j.yexmp.2018.12.005>  
 Right, data from Crossref <http://api.crossref.org/works/10.1016/j.yexmp.2018.12.005>





Experimental and Molecular Pathology  
Volume 106, February 2019, Pages 102-108

## RETRACTED: Ligustrazine promoted hypoxia-treated cell growth by upregulation of miR-135b in human umbilical vein endothelial cells


Shujing Wei<sup>a,b</sup>, Hui Wang<sup>a,b</sup>  

Show more 

 Share  Cite

<https://doi.org/10.1016/j.yexmp.2018.12.005>

Referred to by

Retraction notice to "Ligustrazine promoted hypoxia-treated cell growth by upregulation..."  
 Experimental and Molecular Pathology, Volume 127,  
 August 2022, Pages 104786  
 Shujing Wei, Hui Wang  
 View PDF

```
{
  "status": "ok",
  "message-type": "work",
  "message-version": "1.0.0",
  "message": {
    "indexed": {
      "date-parts": [
        [
          2023,
          4,
          11
        ]
      ],
      "date-time": "2023-04-11T07:10:45Z",
      "timestamp": 1681197045530
    },
    "update-to": [
      {
        "updated": {
          "date-parts": [
            [
              2019,
              2,
              1
            ]
          ],
          "date-time": "2019-02-01T00:00:00Z",
          "timestamp": 1548979200000
        },
        "DOI": "10.1016/j.yexmp.2018.12.005",
        "type": "retraction",
        "label": "Retraction"
      }
    ]
  }
}
```

Sharing hand-validated metadata as well as metadata quality procedures could be helpful in the future. Only public domain data sources can be readily shared; at the time of our study, Crossref and PubMed were the main public domain data sources of retraction data, while Retraction Watch data was licensed. Data availability and licensing might have impacted our findings; for instance, Clarivate, the parent organization of Web of Science, at that time licensed Retraction Watch data for EndNote and presumably could have used it for Web of Science as well.

The state of open data significantly changed after we completed this research: On September 11, 2023, Crossref announced that they had acquired Retraction Watch data and were making it open for an initial 5-year term. At that time, they reported the number of retracted publications as ~14,000 in Crossref, around ~43,000 in Retraction Watch, and ~50,000 in total given the overlap. A longitudinal study should assess whether making Retraction Watch's high-quality, hand-validated data available as open data increases the agreement in retraction indexing over time.

More disagreement was found in items retracted in 2022 and 2023, suggesting that existing data sharing as of April 2023 might have been helping, but might need more frequent updating. Our results suggest significant room for improvement in retraction indexing quality in these

multidisciplinary sources. Fully automatic processes will not be sufficient for creating a comprehensive union list from our current sources, in their current state of data quality.

### **Open science practices**

Code is available at: <https://github.com/infoqualitylab/retraction-indexing-agreement> and archived in Zenodo as <https://doi.org/10.5281/zenodo.8336538>

We have shared data from the Crossref API as:

Lee, Jou; Schneider, Jodi (2023): Crossref data for Assessing the agreement in retraction indexing across 4 multidisciplinary sources: Crossref, Retraction Watch, Scopus, and Web of Science. University of Illinois at Urbana-Champaign. [https://doi.org/10.13012/B2IDB-9099305\\_V1](https://doi.org/10.13012/B2IDB-9099305_V1)

We have shared the stopwords and keywords used to manually identify fields as:

Salami, Malik; Lee, Jou; Schneider, Jodi (2023): Stopwords and keywords for manual field assignment for the STI 2023 paper Assessing the agreement in retraction indexing across 4 multidisciplinary sources: Crossref, Retraction Watch, Scopus, and Web of Science. University of Illinois at Urbana-Champaign. [https://doi.org/10.13012/B2IDB-8847584\\_V2](https://doi.org/10.13012/B2IDB-8847584_V2)

Data for this study is licensed by each source. Only the Crossref API allowed us the right to share the data at the time we collected it. For Retraction Watch Data, we used data available from The Center for Scientific Integrity, the parent nonprofit organization of Retraction Watch, subject to a standard data use agreement (since this work was completed prior to the September 11, 2023 opening of that data). Retracted publications listed in Scopus and Web of Science data can be retrieved from the user interface as shown in Table 1, by database subscribers. Note that checking coverage in Scopus requires specific permission since the Academic Use Case of Scopus API is limited to a single subject area.

### **Acknowledgments**

For feedback, we thank members of NISO Communication of Retractions, Removals, and Expressions of Concern aggregator/end user subgroup. This work would not have been possible without data from the sources used. Thank you to Crossref for providing a public REST API with data that may be used for any purpose. Thank you to Web of Science for API data access. Thank you to the Elsevier ICSR Labs and Scopus API teams for facilitating data access. We particularly acknowledge The Center for Scientific Integrity for free provision of Retraction Watch data for scientometric and data quality research. We appreciate feedback we received from Tilla Edmunds, Yuanxi Fu, Tzu-Kun Hsiao, Rachael Lammey, and Ivan Oranksy and the STI2023 reviewers: Laurent Jégou and 2 anonymous reviewers.

### **Author contributions**

CRedit:

Conceptualization: Jodi Schneider

Data Curation: Heng Zheng and Jodi Schneider

Funding acquisition: Jodi Schneider

Investigation: Jou Lee, Malik Salami, Jodi Schneider, and Heng Zheng

Methodology: Malik Salami and Jodi Schneider  
Project administration: Jodi Schneider  
Resources: Jodi Schneider  
Software: Jou Lee, Malik Salami and Tzu-Kun Hsiao  
Supervision: Jodi Schneider  
Validation: Malik Salami and Heng Zheng  
Visualization: Jou Lee and Malik Salami  
Writing – original draft: Jodi Schneider, Malik Salami, and Heng Zheng  
Writing – review & editing: Jodi Schneider, Malik Salami, and Heng Zheng

### **Competing interests**

JL, HZ, and MS declare no competing interests.

JS declares non-financial associations with Crossref; COPE; International Association of Scientific, Technical and Medical Publishers; the National Information Standards Organization; and the Center for Scientific Integrity (parent organization of Retraction Watch). The National Information Standards Organization is a subawardee on her Alfred P. Sloan Foundation grant G-2022-19409.

### **Funding information**

This project was funded by Alfred P. Sloan Foundation G-2022-19409 Reducing the Inadvertent Spread of Retracted Science II: Research and Development towards the Communication of Retractions, Removals, and Expressions of Concern.

### **References**

- Bakker, C., & Riegelman, A. (2018). Retracted publications in mental health literature: Discovery across bibliographic platforms. *Journal of Librarianship and Scholarly Communication*, 6(1), eP2199. <https://doi.org/10.7710/2162-3309.2199>
- COPE Council. (2019). *Retraction Guidelines*. <https://doi.org/10.24318/cope.2019.1.4>
- Dal-Ré, R., & Ayuso, C. (2020). For how long and with what relevance do genetics articles retracted due to research misconduct remain active in the scientific literature. *Accountability in Research*, 28(5), 1–17. <https://doi.org/10.1080/08989621.2020.1835479>
- Decullier, E., & Maisonneuve, H. (2018). Correcting the literature: Improvement trends seen in contents of retraction notices. *BMC Research Notes*, 11(1), 490. <https://doi.org/10.1186/s13104-018-3576-2>
- Frampton, G., Woods, L., & Scott, D. A. (2021). Inconsistent and incomplete retraction of published research: A cross-sectional study on Covid-19 retractions and recommendations to mitigate risks for research, policy and practice. *PLoS ONE*, 16(10), e0258935. <https://doi.org/10.1371/journal.pone.0258935>
- Genot, E. J., & Olsson, E. J. (2021). The dissemination of scientific fake news: On the ranking of retracted articles in Google. In *The Epistemology of Fake News*. Oxford University Press. <https://doi.org/10.1093/oso/9780198863977.003.0011>



Hendricks, G. (2023, September 11). News: Crossref and Retraction Watch [Website]. *Crossref*. <https://www.crossref.org/blog/news-crossref-and-retraction-watch/>

Hsiao, T.-K., & Schneider, J. (2021). Continued use of retracted papers: Temporal trends in citations and (lack of) awareness of retractions shown in citation contexts in biomedicine. *Quantitative Science Studies*, 2(4), 1144–1169. [https://doi.org/10.1162/qss\\_a\\_00155](https://doi.org/10.1162/qss_a_00155)

Kotzin, S., & Schuyler, P. L. (1989). NLM's practices for handling errata and retractions. *Bulletin of the Medical Library Association*, 77(4), 337–342.

Lex, A., Gehlenborg, N., Strobel, H., Vuilleumot, R., & Pfister, H. (2014). UpSet: Visualization of intersecting sets. *IEEE Transactions on Visualization and Computer Graphics*, 20(12), 1983–1992. <https://doi.org/10.1109/TVCG.2014.2346248>

Malički, M., Utrobičić, A., & Marušić, A. (2019). Correcting duplicate publications: Follow up study of MEDLINE tagged duplications. *Biochemia Medica*, 29(1), 010201. <https://doi.org/10.11613/BM.2019.010201>

Mine, S. (2019). Toward responsible scholarly communication and innovation: A survey of the prevalence of retracted articles on scholarly communication platforms. *Proceedings of the Association for Information Science and Technology*, 56, 738–739. <https://doi.org/10.1002/pr2.155>

Proescholdt, R., & Schneider, J. (2020, October 22). *Retracted papers with inconsistent document type indexing in PubMed, Scopus, and Web of Science [poster]*. METRICS 2020 workshop at ASIS&T 2020. <https://hdl.handle.net/2142/110134>

Schmidt, M. (2018). An analysis of the validity of retraction annotation in PubMed and the Web of Science. *Journal of the Association for Information Science and Technology*, 69(2), 318–328. <https://doi.org/10.1002/asi.23913>

Schneider, J., Ye, D., Hill, A. M., & Whitehorn, A. S. (2020). Continued post-retraction citation of a fraudulent clinical trial report, 11 years after it was retracted for falsifying data. *Scientometrics*, 125(3), 2877–2913. <https://doi.org/10.1007/s11192-020-03631-1>

Snodgrass, G. L., & Pfeifer, M. P. (1992). The characteristics of medical retraction notices. *Bulletin of the Medical Library Association*, 80(4), 328–334.

Suelzer, E. M., Deal, J., Hanus, K., Ruggeri, B. E., & Witkowski, E. (2021). Challenges in identifying the retracted status of an article. *JAMA Network Open*, 4(6), e2115648. <https://doi.org/10.1001/jamanetworkopen.2021.15648>

Wang, P. (2023). Rising of retracted research works and challenges in information systems: Need new features for information retrieval and interactions. In *Proceedings of the 2023 Conference on Human Information Interaction and Retrieval (CHIIR '23)*, 69–82. <https://doi.org/10.1145/3576840.3578281>